Case Study Summary Report

To achieve the goals of the case study the following steps were taken:-

Step 1:- Data Understanding – There were 9240 rows and 37 columns. There were 17 columns which had missing values with 5 features having more than 40% missing values. There was no duplicate value in the data.

Step 2:- Data Cleaning – 1. Features like Asymmetrique Activity Index, Asymmetrique Profile Index, Lead Quality, How did you hear about X Education, Asymmetrique Profile Score, Asymmetrique Activity Score, Tags, Last Notable Activity, Lead Quality were dropped as these were of no use.

2. We dropped the features like Get updates on DM Content, I agree to pay the amount through cheque, Receive More Updates About Our Courses, Magazines, Update me on Supply Chain Content as they were having one unique value.

3. After checking the above data Prospect ID, Lead Number columns were dropped.

4. Country data was heavily skewed as 95% of the data was only of India. Similar to Country, City data was not required for Model building as X-Education is an online platform. Hence these columns were dropped.

5. What matters most to you in choosing a course was skewed, so we deleted the column.

Step 3:- EDA(Exploratory Data Analysis) – EDA was done on the cleaned data to check the condition of it. It was observed that a lot of elements in the categorical variables were of no use. However numerical variables seemed fine. Though outliers in TotalVisits and Page Views Per Visit showed valid values, but these might misclassify the outcomes and consequently create problems while making inferences with the wrong model. So, we capped the TotalVisits and Page Views Per Visit to their 95th percentile.

Step 4:- Dummy Variables Creation – Dummy variables for the various categorical variables were created. Then all the repeated and redundant variables were removed. So, we had 52 columns after it.

Step 5:- Train-Test Split – Splitting of the data was done in the ratio 70:30 for train and test data.

Step 6:- Feature Rescaling – It was done as:-

1. Min Max scaling was done to scale the original numerical variables.

2. A heat map was then plotted in order to check the correlation among the variables.

Step 7:- Model Building – RFE was done to find the top 20 relevant variables. After that the rest of the variables were manually removed depending on their VIF(VIF > 3) and p-values(p-value > 0.05)

Step 8:- Model Evaluation – For model evaluation, firstly a confusion matrix was made and then by using ROC curve the optimum cut-off value was found. This value was then used to find the accuracy, sensitivity and specificity which was found to be around.

On the basis of precision and recall trade off, we got cut-off value = 0.404 .

Step 9:- Final Model – After applying the learnings from the train model to the test model the following metrics were found as :-

1. Accuracy – 80.63%

2. Sensitivity- 82.1%

3. Specificity – 79.67%

Conclusion :- Major indicators that a lead will get converted to a hot lead:

1. Lead Origin_Lead Add Form : A lead sourced from Lead Origin_Lead Add Form is more likely to get converted

2. Occupation_Working Professional :- Working professionals are more likely to get converted.

3. Lead_Source_Welingak website : A lead sourced from Welingak Website is more likely to get converted.

4. Last Activity_SMS Sent :A lead having SMS sent previously are more likely to get converted.

5. Lead Source_Olark Chat :A lead sourced from Olark Chat is more likely to get converted

Major indicators that a lead will NOT get converted to a hot lead:

1. Last_Activity_Olark chat conversation : Customer who had olark chat conversion, are less likely to get converted into hot leads.

2. Lead Ongin_Landmg Page Submission : Customer who hadLead Ongin_Landmg Page Submission, are less likely to get converted into hot leads .

3. Do Not Email :Customer who choose Do Not Email, are less likely to get converted into hot leads .

Recommendations :- The company should use a leads score threshold of 34 to identify "Hot Leads" as at this threshold, Sensitivity Score of the model is around 81% which is as good as CEO's target of 80%.